DOCUMENT RESUME

ED 117 958 FL 007 319

AUTHOR Pinsent, A.
TITLE The Construction and Use of Standardised Tests of Intelligence and Attainment. Pamphlet No. 3.
INSTITUTION Wales Univ., Aberystwyth. Univ. Coll. of Wales.
PUB DATE [60]
NOTE 53p.

EDRS PRICE MF-$0.83 HC-$3.50 Plus Postage
DESCRIPTORS Achievement Tests; Aptitude Tests; *Bilingual Education; Educational Policy; English; Intelligence Quotient; Intelligence Tests; Measurement Instruments; *Secondary Education; *Standardized Tests; *Test Construction; Test Results; Tests; *Welsh
IDENTIFIERS *Wales

ABSTRACT

The British Education Act of 1944 stipulated that instruction and training be offered according to the ages, abilities, and aptitudes of pupils. One specific problem concerned the entry to secondary schools of pupils from a variety of primary schools. The resulting problem of determining the different aptitudes and abilities has been partially solved by the use of standardized tests. This pamphlet is designed to provide a brief introduction to the methods of constructing and using standardized tests, and to discuss special difficulties encountered in the construction and use of standardized tests in Wales, a mixed language area. Specifically discussed are the various kinds of standardized tests, what is meant by standardization, what such tests determine, the choice of tests, comparison of the results of various tests, and the concepts of mental age, attainment ages and quotients. (CLK)

UNIVERSITY COLLEGE OF WALES
ABERYSTWYTH

FACULTY OF EDUCATION

The Construction and Use of Standardised Tests of Intelligence and Attainment

With special reference to the problems of a mixed language area.

A. PINSENT, M.A., B.Sc.

Pamphlet No. 3

FOREWORD

There can be no doubt that changes in the organisation of educational administration subsequent to the Education Act of 1944 will place greater emphasis on educational guidance, and will render more important the use of standardised tests as a means of diagnosis preparatory to remedial treatment.

That being so, it seems desirable that teachers shall be familiar with the principles underlying the construction and use of these tests.

At the present time, there are few accounts available in the region between mathematical and statistical formulae which frighten the layman out of his wits, and easy-going general accounts of tests and testing, which, by omitting all the technical details of standardising, make the process appear to be misleadingly easy.

If standardised tests are to be used in the classrooms in the ordinary course of educational guidance then the teachers who use them should be familiar with such details as are necessary for correct use of the tests and intelligent interpretation of the results.

In particular, so far as Wales is concerned, there does not appear to exist at the moment any account which deals with the special problems of the construction and use of standardised tests in a mixed language area.

This pamphlet has been written in the hope that it may fill the gaps indicated above.

"Intelligence tests are commonly criticised, the most commonly by perons who have little understanding of the way in which they are used. They are blamed for failing to measure things which they are not intended to measure."

MACRAE. *Talents and Temperaments.*

"A testing programme . . . must be planned for a particular purpose and to suit the needs of particular groups of children. The pupils and not the educational institution or system should be the main consideration in framing such a programme. There is always the danger that extensive testing, particularly where results are not used and interpreted effectively, can lead to rigidity and even sterility in teaching because measurements tend to assume some value in themselves when, in fact, they have only relative value in the light of the action which follows."

"On the other hand teaching without use of carefully planned testing may result in a good deal of misplaced effort on the part of teachers, and failure and frustration on the part of pupils. Not a little of this loss of achievement and frustration on the part of both teachers and taught could be avoided if objective measures of appraisal, diagnosis and checking were used. A knowledge of *standardised tests* of mental ability and achievement should be part of an adequately trained teacher's classroom equipment.

SCHONELL. *Diagnostic and Attainment Testing.*

"Tests have their limitations as well as their values and one should know what points to observe in order to avoid pitfalls of over-evaluation or incorrect interpretation . . . *One should not place too much reliance on an isolated test finding.*"

SCHONELL. *Diagnostic and Attainment Testing.*

CONTENTS

Published by the Faculty of Education, University College of Wales, Aberystwyth, and printed by Gee & Son, Ltd., Denbigh.

1

WHAT THIS PAMPHLET IS ABOUT

Dissatisfaction with Scholastic Examinations

Scholastic examinations have had a very long history. They have been used by people as far apart in time and conditions as the Ancient Chinese and Modern British.

Since 1902 when County Councils in England and Wales were made responsible for secondary education, admission to grammar schools at age eleven-plus has depended increasingly on competitive scholastic examinations. At the same time access to grammar schools has become, economically and socially, increasingly important.

It is not surprising therefore, that scholastic examinations have become very much a matter of public concern. They have been subject to critical scrutiny by educationists, psychologists and statisticians with, in some cases, rather startling results.

If examinations are to be used for selecting candidates for higher education it is essential that they shall be fair and reliable. The examination of examinations, as it has been called, has aroused serious doubts about the reliability of the academic essay of type of examination particularly for *predicting* future educational progress and ultimate level of scholastic attainment.

These doubts have stimulated a more careful analysis of the examination procedure itself, as well as of the results it was supposed to achieve.

For our purpose, perhaps the gravest defect of the traditional scholastic examinations was their failure to distinguish, reliably, between present attainment and probable future performance. Properly conducted, these examinations do indicate that the successful candidates have already attained a certain level of skill in writing answers in the form of essays, and have acquired certain types of usable information. However, follow-up studies of grammar-entrance examinations at age eleven-plus have shown, beyond doubt, that as instruments for *predicting* future academic successes they are by no means reliable. Yet when used as competitive tests for grammar-school selection or university entrance, the *predictive* function of the examinations is more important than

the measure of present attainment. What matters in those cases is not merely how much the candidate has been coached to learn up to the present moment, but rather, how rapid will be his progress in the next few years and how high his ultimate achievement.

Comparative studies have revealed serious defects in the examination procedure itself—variations in the standards and methods of marking, for example. In addition, it has been suggested that success depended to a significant extent on conditions which were not at all equal for all candidates. Good school conditions, efficient teaching, regular attendance, unbroken schooling, good health, good homes, a good memory, speed of writing, all favour success. On the other hand, bad teaching, poor school conditions, ill-health, frequent changes of school, poor homes, emotional maladjustments all hinder a candidate from rising to the level which his real intellectual ability would indicate.

Search for more Reliable Tests.

Consequently, since the beginning of the present century, when the *selective* indications of scholastic examinations increased in economic as well as academic importance, we have seen a persistent search for more reliable tests of intellectual aptitude and future scholastic attainment.

This search has led, among other results, to the adoption and refinement of objective standardised tests as supplements of, or alternatives to traditional scholastic examinations. Professor Sir Godfrey Thomson, one of the prime movers in this search said in an address to the National Foundation for Educational Research that he felt that he had a moral duty to do everything possible to improve methods of discovering intelligent children who might be overlooked, and guiding them into forms of higher education likely both to make them happier in their lot and useful to a society and civilisation which needs them.*

Misgivings about Standardised Tests.

However, the use of standardised tests, of "intelligence" particularly, to select candidate at age eleven-plus for admission to grammar schools, has aroused as much criticism and opposition in the lay public as the unreliability of essay-type examinations aroused in the experts. The "intelligence" tests are viewed with marked suspicion which has been exploited by "sob-stuff" writers in certain newspapers and magazines, usually people with little or no real knowledge of the methods by which these tests are prepared

*Bulletin of the National Foundation for Educational Research. No. 2. November, 1953.

and validated. Against this, there is now much experimental evidence that, although by no means perfect, correctly standardised tests of "intelligence" are more reliable *predictors* of general success than traditional essay-type examinations. Suspicion in the minds of parents and the general public has arisen mainly from the fact that these tests would appear to have been used solely for the purpose of excluding all but a small proportion of children of eleven-plus from the secondary grammar schools.

This, however, is not a valid criticism of the tests. It is rather an indication of the unsatisfactory nature of our educational system. The effective answer is not to abolish standardised tests but to provide more, and more adequate, secondary school accommodation.

Reasons Why Standardised Tests are Likely to be Increasingly Used.

This unfortunate association of tests of "intelligence" with selection for grammar school places has diverted attention from other important functions in educational practice which standardised tests can fulfil without incurring the suspicions we have noted. There are indications at the present time that even if examinations for grammar school selection should be abolished entirely, the need for standardised tests of attainment as well as "intelligence" will become increasingly important and increasingly wide-spread in the future.

Implementing the Education Act of 1944.

It is not yet realised explicitly even in educational circles, how revolutionary in relation to traditional English notions was the Education Act of 1944. This enacts that each Local Education Authority must provide such educational facilities as will afford "for all pupils such opportunities for education offering such variety of instruction and training as may be desirable in view of their different ages, abilities, and aptitudes and of the different periods for which they may be expected to remain at school, including practical instruction and training appropriate to their respective needs."

This provision of *varied instruction and training* in view of different ages, abilities and aptitudes implies two processes—(a) providing school buildings and material facilities for the varied types of training, and (b) devising methods of discovering what different abilities and aptitudes do, in fact, exist and what types of training are most suitable for their nurture, or, in other words, organising processes of educational guidance. How can item (b) be accomplished reliably?

Human institutions tend to be self-perpetuating. They persist with a perverse tenacity long after the conditions which brought them into existence originally have ceased to be important. This is certainly true of English education. That has been organised for centuries on the assumptions that the only type of education which mattered was provided by the classical grammar schools; and the only type of ability and aptitude worth serious nurture was that which thrived on the classical grammar-school curriculum. It was most unfortunate that just before the passing of the 1944 Education Act these traditional attitudes were embodied in the psychological myths of the Norwood Report which was regarded by people already steeped in the tradition as a form of Holy Writ.

The challenge of the 1944 Act needs to be taken up. Are there, in fact, abilities and aptitudes other than the linguistic, abstract, intellectual? If so, what are they; what is their educational importance for both the community and the individual in the modern world; at what age do they appear: and how can they be detected and trained to full maturity? This is the problem of educational guidance in which standardised tests are likely to play an increasingly important part.

The Situation in Comprehensive, Multilateral and Bilateral Secondary Schools.

Some Education Authorities are making experiments with comprehensive or 'multilateral' or 'bilateral' secondary schools. In these cases, all the pupils in an administrative area who are above the level of the just-not-certifiably-feeble-minded go into the same secondary school. This practice obviates the necessity for a selection examination at eleven-plus, but the staff of any comprehensive secondary school will still be faced with the need to sort out pupils into kinds and grades of ability and attainment and to adapt both curriculum and methods of teaching to pupils' varying needs. National welfare as well as individual educational progress demands, nowadays, that the best possible use shall be made of all our available ability at whatever level it is manifested. Moreover it is absurd to teach the highly gifted in the same way and at the same pace as those of average or lower than average mental aptitudes, apart altogther from the adequate treatment of different kinds of aptitudes.

Again, it is becoming increasingly popular to suggest that any sorting process necessary on entry to secondary education can best be done on the basis of cumulative records of the progress of individual pupils throughout their primary school careers. However, this alternative introduces its own particular difficulties. Pupils entering any one secondary school will be recruited from several different primary schools. How then can the estimates of aptitudes

and attainment made by different teachers in different schools be *compared fairly*? There is no lack of objective evidence that the judgments of different teachers even in the same school vary significantly. Therefore, if pupils on entry to secondary schools are to be sorted out reliably on the basis of records compiled in different primary schools there must be some guarantee that the records in question are all expressed in terms of one standard scale. As we shall see later, a standardised test of "intelligence" or attainment is itself a standard scale. Therefore all records of aptitude or attainment obtained from the *same* standardised test correctly administered according to the prescribed instructions can be compared fairly, even though the pupils concerned come from different schools.

Backwardness and Educational Retardation.

Further, the introduction of secondary education for all according to the Act of 1944 has revealed to the horrified gaze of secondary school teachers the extent and difficulties of the problems of backwardness and educational retardation.

Previously, when admission to secondary schools was determined largely by success in a scholastic examination, secondary school teachers could reasonably expect that pupils of eleven-plus should have reached an accepted minimum standard of attainment in language, reading and arithmetic.

This expectation is no longer reasonable. When all pupils of eleven-plus proceed automatically to secondary schools, then the secondary schools must perforce accept a wide range of aptitude and attainment in their intake. At age eleven-plus there may be a range of attainment as wide as from below 7 to 15 years. That being so, it is useless for secondary school teachers to accuse their primary school colleagues of inefficiency, or worse. Instead of demanding a minimum level of scholastic attainment according to traditional scholastic standards the only demand that can now be made legitimately by the secondary school teachers is, that the pupils from the primary schools shall have reached a level of attainment in the three R's *on a par with their educable capacity*. Thus, if a pupil of 11 years has a mental age of only 9 years, then if his attainment ages in language, reading and arithmetic are 9 years, that pupil has been efficientiy taught, although to the secondary school teacher he may appear to be seriously backward. Secondary schools in the new dispensation will have to learn not to expect the impossible.

Moreover they themselves will have to deal efficiently with the problems of the backward children. It is essential in the treatment of a backward child to discover (a) whether the child is merely retarded or (b) whether he is backward because of innate dullness.

In either case, successful treatment will depend on adapting the level of difficulty of his work to that of his real mental capacity.

The most reliable way of estimating the mental and attainment ages of a group of pupils is by means of well-standardised tests. They are not only more reliable predictors of general educational capacity and future academic progress than the traditional examinations; they are also much more sensitive indicators of scholastic defect. An actual example will illustrate the value of standardised tests for diagnostic purposes. A group of 111 pupils, aged eleven-plus, the yearly intake of a secondary modern school, was tested in vocabulary and reading comprehension with the following results—*

SCHONELL VOCABULARY TEST.

Vocabulary Age	No. of Cases	Vocabulary Age	No. of Cases
15	1	10	22
14	3	9	16
13	2	8	8
12 (normal)	5	7 or less	33
11	21		
		Total	111

SCHONELL TEST OF READING COMPREHENSION.

Reading Age	No. of Cases	Reading Age	No. of Cases
15	0	10	20
14	0	9	25
13	2	8	21
12 (normal)	1	7 or less	30
11	12		
		Total	111

Here we find a secondary school intake all at the same chronological age with a range of vocabulary ages from less than 7 up to 15 years, and of reading ages from less than 7 up to 13 years. Obviously this intake must be taught by methods and exercises appropriate to their attainment ages. It is possible, nowadays, when the attainment ages have been discovered by the use of standardised tests, to consult schedules which indicate appropriate books

* For these data I am indebted to Mr. E. S. Thomas, Pembroke Dock.

and exercises for given vocabulary and reading ages.† This example will serve to show how much more definite and therefore practically useful are the data revealed by the standardised test than impressions gained by casual observation.

New conditions, even in educational administration and teaching practice, demand new methods. The need for differentiation of curricula in order to adapt them to the needs of pupils of different levels of intellectual capacity and maturity, particularly at the secondary stage, is likely to demand the application of standardised tests to an increasing extent in the not-too-distant future. It seems desirable, therefore, that some knowledge of the construction and use of these tests should become an established part of every teacher's training.

Special Problems in Wales

So far as differentiation of schools and curricula and the special problems of secondary education are concerned, administrators and teachers in Wales are faced with difficulties of the same type as their colleagues in England. However, in Wales there are special problems in connection with the mixture of languages and the implementation of a bilingual education policy. As we shall see later this policy introduces problems the solution of which would be helped very much by the use of standardised tests. At the same time, the construction and standardisation of tests in Welsh, and the use of standardised tests whether Welsh or English involves difficulties peculiar to Wales.

The object of this pamphlet is to provide a brief introduction to modern methods of constructing and using standardised tests and to discuss the special difficulties involved in this work in Wales. The topics will be treated here only in sufficient detail to give teachers and educational administrators some insight into the nature, construction and use of these tests for practical purposes. Readers who are interested can follow the topics into more detail in the references indicated in the bibliography.

†See, for example, Schonell: *Diagnostic and Attainment Testing* (p. 162).

2

WHAT ARE STANDARDISED TESTS?

COMPARISON WITH SCHOLASTIC EXAMINATIONS

An accepted canon of good teaching is to proceed from the known to the unknown. We can, with profit, use the principle in an approach to the discussion of standardised tests and begin with what all teachers are only too familiar, the traditional examination.

Any examination represents an *attempt to take samples* of a candidate's total knowledge and skill. From the results of the four, five, six, seven, or eight sample questions which the traditional examination paper contains, the examiner judges whether or not the whole of the candidate's knowledge or skill is satisfactory.

The process of examining is exactly similar in form to that used by a buyer of some commodity in bulk. It is impossible to view, in the time available, all the articles to be bought. Therefore the buyer takes out samples here and there from the bulk, and makes his decision about the probable quality of the whole consignment on the quality of the small samples. Obviously, the accuracy of his judgment will depend on the skill or the luck with which he chooses his samples, and on the number of samples he takes. The fewer the samples the more will luck and biassed judgment interfere with the verdict.

Thus the first objection which can be urged against the traditional examination is that it contained too few questions. It did not sample the candidate's total knowledge or ability fairly. The element of chance or 'luck' played too great a part in the verdict.

In the second place, in large scale examinations such as the School Leaving Certificate,* the answers were marked by many different examiners, each with his own particular (and often peculiar) methods of marking and standards of assessment. (See Fig. 1).

Further, the relative difficulty of each question, and its value in marks was decided by the personal subjective judgment of whoever set the questions without reference to the actual degree of difficulty the questions might present to the candidates tested.

These difficulties might not have serious consequences if the examinations were used in school as routine tests by teachers with personal knowledge of the candidates. They were most serious, however, when the examinations were used for competitive pur-

* Now General Certificate of Education.

poses, e.g., selection for grammar school places, or admission to a university, or state scholarships.

Compared with traditional scholastic examinations, standardised tests contain a large number of questions, each requiring a short, unequivocal answer. In this way they sample the testee's abilities and attainments much more thoroughly, thus reducing the element of chance.

Writing is reduced to a minimum.

Administration of the tests is standardised. Each test must be given strictly according to carefully prescribed instructions. These ensure, as far as possible, uniform conditions for all candidates.

Marking the answers is standardised. Exact directions for marking, and the mark values of each answer are prescribed in the manual of instructions.

Thus, the tests may be given and marked without varying the conditions and standards of marking by any person capable of reading and understanding the instructions, and honest enough to obey them exactly.

Finally, the standardised test is always tried out sometimes on several occasions before it is published *on a representative population of pupils as similar as possible to those for whom the test is intended* in order to discover the relative difficulty and power of discrimination of the test questions, and their general suitability for their purpose.

This process of preliminary trial and what is called item-analysis will be described in more detail later.

3

COMMON TYPES OF STANDARDISED TESTS

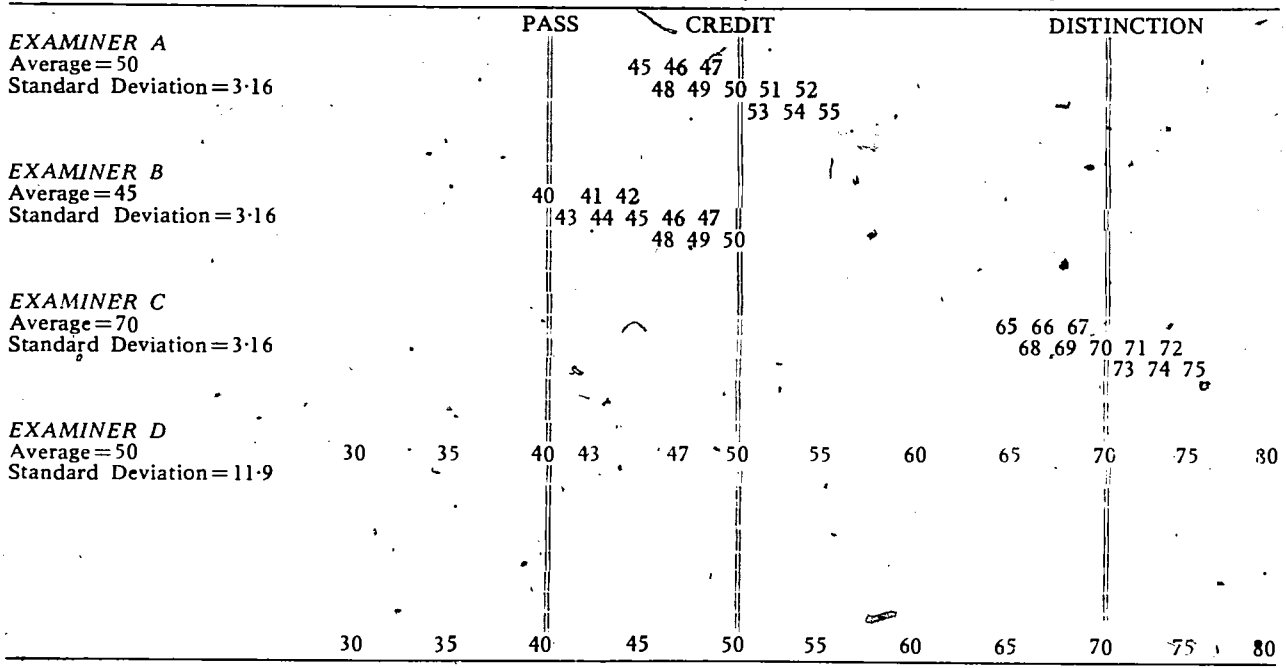
It is convenient at this stage to indicate the principle types of tests in common use since the form of the test determines the details of the processes of standardisation.

The first standardised tests were designed to be given to children individually. This, however, is a lengthy process. A Binet test may need anything from a half to one and a half hours to give. As the used of standardised tests increased, other types were devised for simultaneous administration to groups. The American Army Tests during the 1914-19 war were given to as many as five hundred men at a time.

In addition to the distinction between individual and group tests,

FIGURE I

To illustrate types of variation between different examiners in essay-type examinations with respect to average mark and scatter of marks.



IDEAL MARK SCALE

Each examiner is assumed to have marked eleven scripts and to have awarded the marks shown opposite to his designation. Types of marking illustrated above can be found in actual practice.

(For definition of standard deviation, see page 18).

the tests vary also according to the purpose for which they are intended.

Some are tests of "intelligence"—more accurately, of intellectual capacity or general educational aptitude. In these the questions require, predominantly, the exercise of powers of observation, reasoning and ingenuity, i.e. general abilities common to many types of intellectual and practical activities. These tests are used to indicate a pupil's probable rate of intellectual development, and the upper level of attainment which, given *favourable conditions*, he may reasonably be expected to achieve in the not too distant future.

On the other hand, some standardised tests are intended to discover a pupil's present level of scholastic attainment in, for example, mechanical arithmetic, problem arithmetic, spelling, word-recognition, vocabulary, reading-comprehension, sentence-structure.

The use of verbal tests of educable capacity presents special difficulties in cases of backward readers, and in mixed language areas. The latter will be discussed in more detail later. For these reasons, non-verbal and performance tests have been devised.

Non-verbal tests are "pencil and paper" tests using visual forms of test material in which the use of words is reduced to a minimum.

In performance tests, the pupils tested are required to perform some practical activity such as fitting together jig-saw patterns, or threading mazes.

Non-verbal and performance tests have the advantage that the instructions for procedure can be given equally well in any language. Dependence on particular word-habits is reduced to a minimum even though it may not be eliminated completely.

This classification of standardised tests can be represented in summary form as follows—

Individual tests (e.g. Terman-Binet or Terman-Merrill) to be used for testing one individual at a time.

Group tests, to be used for testing whole classes at the same time.

Both the above types may include—

(a) Tests of "intelligence," that is of general educational capacity.
  (i) verbal
  (ii) non-verbal
  (iii) performance

(b) Tests of attainment in particular scholastic subjects.

Certain tests for special abilities (e.g. musical, mechanical, artistic, clerical) are now available. These are more important for vocational selection and guidance and need not concern us here.

Two further distinctions should be noted. They are important

for our purpose for two reasons: (a) they imply differences in methods of standardisation, (b) they need to be taken into account when choosing tests for particular purposes and interpreting the results.

No test, however accurately standardised, will give absolute estimates. The test results must always be interpreted in relation to the method of standardisation; the population used for standardisation; the purpose of the test; and the population to be tested. Standardised tests are measuring devices. As such they must be used *only in the conditions and for the objectives for which they have been constructed*. Only too often, tests are used for some purpose for which they are not appropriate. Then if the results differ from those which were expected all standardised tests are damned without qualification. It is just as reasonable to quarrel with a yard stick because it will not measure pounds or pints.

The first distinction is between tests which cover extended as against restricted age-ranges.

The earliest tests of the Binet type covered an age-range between five and fourteen years. The latest Terman-Merrill revision of the Binet test extends from two to eighteen years. Other tests have ranges of six to eleven years; seven to fourteen years, for example.

However, the practical use of these tests revealed statistical and other difficulties, particularly at the extreme upper and lower ends of the scale. Partly for this reason, and partly on account of the increasing use of standardised tests for selecting candidates at age eleven-plus for grammar school entrance, tests were standardised for particular restricted age-ranges, e.g. 10 to 12 years—the limits usually prescribed by Local Education Authorities for the grammar school selection tests.

The restricted age-range tests have a further advantage. They are more sensitive to small differences between the various candidates within the group. This is a most important consideration in the selection process at age eleven-plus.

The second distinction is that between age-scales and point-scales as they may be called.

In the age-scales, each test item is evaluated directly in terms of mental or scholastic age.

In the point scales, each test item is given a prescribed mark; the total marks (or points) gained by the testee on the whole test are tabulated, and these totals are then equated with mental or scholastic ages.

The meaning of the terms mental age and scholastic or attainment age will be explained in more detail later. For the present it is sufficient to know that mental age is applied to scores on a test of "intelligence" of the Binet type; scholastic or attainment ages correspond to scores on tests of scholastic attainment. To illustrate methods of standardisation, the discussion will be restricted to tests of intelligence. This will avoid tiresome repetition.

HOW ARE STANDARDISED TESTS STANDARDISED?

A standardised test is, essentially, a scale of mental or scholastic measurement graduated in convenient units of intellectual or scholastic development.

Standardisation involves several processes, e.g.—

(a) discovering *by actual trial*, the relative difficulty and discriminating power of each test item;
(b) arranging the test items in an order of relative difficulty;
(c) devising convenient units of mental or scholastic development and attainment;
(d) arranging the units to form a reliable scale.

In actual practice, standardising a modern test to an acceptable degree of reliability is a complicated and highly-skilled process for which special training and statistical knowledge are essential. However, we are interested here primarily in the general principles underlying the construction and use of these tests. The principles themselves are not really difficult to understand.

A Statistical Digression.

It is desirable to digress at this point to consider three common statistical terms—

(a) Arithmetic Mean or Average.

Table I illustrates a mark-list for a class of 20 pupils in a test with a maximum mark of 20. The marks gained by each pupil A, B, C, D etc., are shown in Column I.

TABLE I

To illustrate calculation of arithmetic mean and standard deviation.

Pupil	Col. I Mark	Col. II Deviation from Av.	Col. III Deviation Squared		Col. I Mark	Col. II Deviation from Av.	Col. III Deviation Squared
A	17	+7	49	L	16	+6	36
B	12	+2	4	M	14	+4	16
C	3	−7	49	N	15	+5	25
D	6	−4	16	O	4	−6	36
E	9	−1	1	P	10	0	0
F	10	0	0	Q	13	+3	9
G	11	+1	1	R	10	0	0
H	7	−3	9	S	10	0	0
I	9	−1	1	T	5	−5	25
K	8	−2	4	V	11	+1	1
			Totals		200		282

The *average* or *arithmetic mean* is the sum of all the marks divided by the number of cases. In the example shown, the total marks gained by all candidates amount to 200. There are 20 candidates. Thus the average mark for this class in this test is $200 \div 20$, i.e. 10.

The average represents what may be called the central tendency or general level of the group. If there were a bigger proportion of clever pupils in the class, the average would be accordingly higher; if more dull pupils, the average would be lower.

(b) Standard Deviation.

However, in considering the distribution of marks within a group it is not sufficient to know the average or central tendency only. We need to know also the extent to which the marks are spread or scattered along the mark scale. The usual measure of spread, or scatter of a distribution is the *standard deviation*.

This is calculated as follows—

The deviation or difference of each score from the average is computed (See Column II, Table I);

Each deviation is then squared (Column III);

The squared deviations are added together;

The sum of the squared deviations is divided by the number of cases to get the *mean squared deviation*.

Finally we find the square root of the mean squared deviation which gives the required standard deviation.

For those who prefer a formula—

$$\text{Standard deviation} = \sqrt{\frac{\text{sum of squared deviations from the mean}}{\text{number of cases.}}}$$

In our example:—

$$\text{Standard deviation} = \sqrt{\frac{282}{20}} = \sqrt{14.1} = 3.75.$$

EFFECT ON MARK LISTS OF DIFFERENCES IN EXAMINERS' AVERAGES AND STANDARD DEVIATIONS.

We can now understand more clearly how differences can arise between different examiners marking the same batch of scripts (See Fig. I).

Examiners differ with respect to their customary average mark; some are more generous, others more exacting. They also differ in the spread of the marks along the mark scale between full marks and no marks. In some cases, candidates are bunched close together

in the upper, middle, or lower ranges of the mark scale. In others, the marks are splashed quite freely throughout the whole range. These tendencies are found to be characteristic of individual examiners and teachers, almost as characteristic as their signatures.

Obviously then, the idiocyncrasy of an examiner may be a matter of crucial importance in the case of selection for secondary schools, or when distinctions and failures are in question. Candidates whose scripts are marked by an examiner who has a generous average, and who bunches his marks close together in the upper range of the mark list will have a decided, and, probably, quite unmerited advantage.

In competitive examinations, therefore, it is imperative that all examiners' mark lists shall be transformed to one standard mark scale with a constant average and constant scatter (standard deviation). Both statistics must be controlled at the same time since even if two examiners' average marks are the same, their scatter may be widely different.

(c) The Normal Distribution of Scores.

This need for one standard mark scale into which the marks of individual examiners may be transformed brings us to the conception of a normal distribution.

The question arises immediately, which particular standard scale shall be adopted for the purpose of scaling marks in examinations and tests? Will any standard scale be satisfactory, or is there one with special advantages?

This question has been answered in practice, by noting how "intelligence" or scholastic attainments are actually distributed in a large unselected population of pupils. By 'unselected' is meant that no bias has been used in choosing members of the group—all possible variations are represented *in due proportion*.

Many hundreds of thousands of pupils have been tested in various countries for "intelligence" and scholastic attainment. In all cases, where the samples tested have been large, and in the absence of special bias in selection, it has been found that the results of the tests arrange themselves not quite exactly in a normal distribution, but in very close approximation to it.

The exact details of a normal distribution need not concern us here. They are described and illustrated in any primer of statistics.* It is important for our purpose to note that when large, unselected populations of pupils are given the same tests, their marks, and by implication their "intelligence" and attainments, do approximate closely to this type of distribution. Further, when the average and standard deviation of a normal distribution are known the proportions of candidates in a normally distributed population which

* See for example, a clear and simple exposition in Glassey and Weeks' *The Educational Development of Children.*

may be expected to appear within any part of the mark scale, can be calculated. This is important since it implies that the known characteristics of the normal type of distribution can be used in order to check the trials of a standardised test for bias and other irregularities.

For example, if an "intelligence" test is given to a sufficiently large and unselected population, and if the average I.Q. is 100, and the standard deviation of the I.Q. scores is 15, we may expect the percentages of cases in Column I of Table II at each I.Q. level. If the average is 100 and standard deviation 16·5 instead of 15, the percentages will approximate to those given in Column II.

Conversely, by comparing any distribution of the marks with a normal distribution having the same mean and standard deviation we can estimate the likelihood that the marking is biased or the population specially selected.

TABLE II

Showing the percentage of cases which may be expected at given levels of Intelligence Quotient if the distribution is normal and the standard deviation known. The average in both cases is 100.

	I	II
I.Q. Level	Proportion to be expected if standard deviation is 15	Proportion to be expected if standard deviation is 16.5
131 and over	2.2%	3.6%
121 to 130	6.8%	7.4%
111 to 120	16.0%	16.0%
101 to 110	25.0%	23.0%
91 to 100	25.0%	23.0%
81 to 90	16.0%	16.0%
71 to 80	6.8%	7.4%
70 and less	2.2%	3.6%

Note the difference in the proportions due to increase in the extent of the scatter or dispersion. The larger the standard deviation the higher the proportions of cases at either end of the distribution.

Thus, the normal distribution is accepted by most test-constructors as a convenient basis for an ideal or standard mark-scale. Psychologists believe that "intelligence" and scholastic attainments are, in fact, normally distributed.* Therefore, most standardised tests are deliberately arranged so that if administered to large unselected populations, e.g. all the pupils in a county area, the results expressed as intelligence or attainment quotients will fall

* This belief has never been absolutely proved. It is a convention accepted as a working hypothesis based upon circumstantial evidence.

into a normal distribution. If at the first trials the distribution is not normal the test is readjusted until the scores do approximate sufficiently closely to the normal distribution.

STANDARDISING A TEST FOR A WIDE AGE RANGE.

The best-known test of this type is the Stanford-Binet Scale for measuring "intelligence."

This test was first devised by a French psychologist Binet at the beginning of this century. After his death Burt in London and Terman at the Stanford University in California extended and improved the test. Terman and Merrill issued a revised and restandardised American Edition in 1937.†

The main principles of test-construction and standardisation can be followed by reference to this Binet Scale.

ITEM ANALYSIS.

In the first place a large number of short questions is collected, several times as many as will be needed in the finished test.

Since the Binet test is intended for an age range from 2 years to 18 years the list of problems must be arranged in an order of increasing difficulty to correspond with the growth of mental capacity and experience from infancy to maturity.

In addition to arranging the test-items in ascending order of difficulty. it is also necessary to find out whether they *separate the various age-levels sufficiently clearly* and by approximately regular intervals. It is obvious that if a test item is answered by as many four-year-olds as five-year-olds it will not discriminate sufficiently clearly between the five-year-old and four-year-old levels of "intelligence."

The process of discovering the degree of difficulty and the power of each test item to discriminate between levels of ability is known as item analysis. It is an essential feature of all test-construction. It is in this respect. as well as in using a much larger number of questions that the standardised test differs most significantly from the orthodox essay or 'quiz' type of examination.

How can these qualities of the test items be discovered?

The questions are tried out by experiment on a large and representative population of pupils *characteristic of the region in which the test will afterwards be used.* The percentages of pupils at each age who answer each test item correctly are tabulated. This process is illustrated in Table III. The degree of difficulty of any test item is estimated by the percentage of pupils who answer it correctly. If 100% give correct answers the item is too easy for

† See Terman L.M. and Merrill, M.A., *Measuring Intelligence*. Harrap.

that age-group; if none answer it correctly it is too difficult. By taking averages along the rows of percentages a total order of difficulty can be estimated (see last column, Table III). These averages from item I downwards should decrease by approximately equal amounts. In particular, there should be no reversals in the descending order.

The first trial of a list of test-items usually reveals various anomalies. Some items are ambiguously worded. Others are not suitable for children of that particular area.* These items are reworded or discarded. Some items will be incorrectly placed in the order of difficulty. It is quite impossible to discover how difficult a particular item is until it has been tried out on a sample of pupils. Reversals in the order of difficulty may be removed by rearranging the order of test-items or by suitable rewording.

TABLE III

The numbers in this table are hypothetical, for purposes of illustration only.

Test Item	Ages 5	6	7	8	9	10	11	12	Average
1	50	70	90	99	99	100	100	100	88.5
2	45	65	85	95	97	98	100	100	85.6
3	35	55	75	90	95	97	98	100	80.6
4	25	40	60	85	90	93	95	99	73.4
5	15	25	45	70	80	85	90	95	63.2
6	5	15	30	55	70	80	85	90	53.7
7	0	5	15	40	60	75	80	85	45.0
Etc.	—	—	—	—	—	—	—	—	—

GRADUATING THE SCALE. WHAT UNITS SHOULD BE USED?

Binet's aim was to construct a scale which would measure in readily understandable units whether any pupil was educationally backward, normal, or advanced *for his age*. He decided therefore to graduate his scale in units of achievement-for-age. Thus, if a child of 10 years could pass the test items answered correctly by the *average* child of 12 years then he might be said to have a mental (or educational) age of 12. He would be two years of mental development ahead of his own chronological age. If the ten-year-old pupil passed the tests only up to those correctly answered by the *average* ten-year-old he would be a normal or average pupil for his age. If he succeeded only up to the tests passed by the *average* seven-year-old pupil his mental age would be seven and he would be, therefore, three years retarded.

* e.g. it would be useless to ask English children questions about dollars, 'quarter', dimes, and cents.

This notion of a standard of achievement corresponding to an average age, and a scale in units of yearly progress is easy to understand and convenient to use. It had been familiar to teachers in elementary schools long before Binet adopted it for his scale. It was implied by the Revised Code for Elementary Education issued by the Committee of Council in 1862. In that year elementary school teachers were, in modern factory terminology, put on to "piece-work" and paid by results. At once it became necessary to devise a method of measuring the results. This was done by dividing the period of school life from 7 years to the leaving age into yearly increments called "standards." Then for each standard a syllabus in the three R's was prescribed and rigidly maintained. Pupils were transferred from the Infants departments to the upper school at age 7 and the normal pupil was promoted to a higher standard once per year.

Thus each year of chronological age corresponded throughout the country to one year of scholastic attainment. On this basis, by comparing the chronological age of a pupil with his "standard," his scholastic status could be assessed immediately. Thus the normal 9-to-10 year old pupil would be in Standard III. If he were in Standard I he would be retarded scholastically by two standards or years; if in Standard IV he would be one standard unit ahead.

Moreover, since the standard syllabuses were prescribed by the central authority in Whitehall for the whole country, it was easy to compare results in different schools. Also, when a pupil moved from one school to another it was equally easy to allocate him to the standard appropriate for his age and ability.

This notion of *achievement for age*, is still used in official documents relating to educational subnormality. The Ministry of Education Pamphlet No. 5 states (p. 19)—

"No child should be classed as educationally sub-normal unless he is retarded in school work, but some agreement should be arrived at on the degree of retardation that would justify special educational treatment . . . It is suggested that a large body of opinion would be found to favour giving special educational treatment if a child is so retarded that his 'standard of' work is below that *achieved by average children* 20% *younger than he is* . . . All degrees of retardation may be found among educable children from the minimum indicated above to a maximum which may be as much as 50% where the child can do school work only as difficult as that done by *average children half his age.*"*

In principle, the Binet scale is the same as the old elementary school standard scale. However, the great merit of the Binet scale lies in the fact that the "standards" of average achievement-for-

* Italics mine — 20% younger is equivalent to a mental ratio of 80/100 i.e., 80 I.Q. The maximum retardation above is 50/100 or 50 I.Q. in terms of the Binet-type scales.

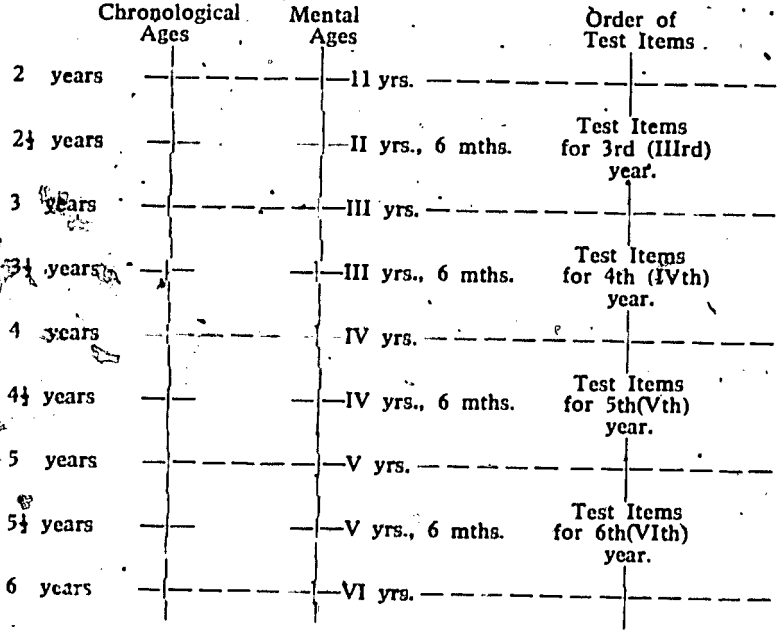
age are determined not by what H.M. Inspectors deem it desirable for pupils to achieve, but by what is actually achieved by a representative sample of pupils at a specified age. The levels of difficulty are fixed for the modern standardised test *by experiment* not by inspection.

Having adopted these units of achievement-for-age it now remains to calibrate the mental scale correctly. This aspect of the problem has aroused more than a little controversy.

For an achievement-for-age scale to work satisfactorily in practice it must be calibrated in such a way that the *average or normal pupil will always score a mental age equal to his chronological age* e.g. the average seven-year-old will achieve a mental age of VII, and similiarly for each age throughout the range covered by the test.

FIGURE II

To illustrate the convention used for scaling Mental Ages corresponding to Chronological Ages.



The problem now is to allocate each test item found to be satisfactory by the item analysis correctly to its appropriate year.

Burt's solution of this problem is as follows.* (The argument can be followed by reference to Figure II). We have a chronological age-scale given by the calendar ages of the pupils. We need a corresponding mental age-scale made up of the test-items arranged in corresponding years of mental development or achievement. It is convenient to use Arabic numerals for chronological age and Roman numerals for mental age.

The pupils are grouped according to their ages at last birthday. Thus, in a large representative sample arranged in order of ages *the average* 4½-year-old pupil will be mid-way in the group who are 4-but-not-yet-5 years and so on for the other age-groups.

Now imagine that all the test items have actually been allocated correctly on the mental-age scale. The items for year V must be suitable for the pupils in the group 4-but-not-yet-5 years in order to give the average child of exactly 5 years a mental age of V. Such a child if correctly measured should pass all the questions appropriate for the Vth year. This may seem somewhat confusing but it resembles the convention we use in calling the years from 1700 to 1799 the 18th century.

Then it follows that to place the *average* 4½-year-old pupil correctly at IV years 6 months on the mental age-scale, approximately *half* the 4-but-not-yet-5-year-old group must pass all the tests for the Vth year.

Hence in standardising a mental age-scale of the Binet type *a test item is correctly allocated to a given mental age if it is passed by 50% of the group nominally one year below that age.* E.g. a test item will be correctly placed for year V if it is passed by 50% of the 4-but-not-yet-5 year-old group.

Perhaps a clearer example of this principle is given in a report on the standardisation of a graded word-reading test by P. E. Vernon.* The word "threaten" was read correctly by the following percentages of pupils—

Age	6 but not 7	7 but not 8	8 but not 9	9 but not 10	10 but not 11	11 but not 12	12 but not 13
%	6	14	29	49	67	84	95

It was, therefore, assigned to a reading age of IX years 6 months. To be assigned a reading age of X years exactly a word would need to be read by 50% of the pupils aged 9½ but not yet 10½, and so on.

Thus, on the basis of the item-analysis unsatisfactory test items are discarded, others are reworded, and the list is rearranged until

* See *Mental and Scholastic Tests.* p. 152.

* See *The Standardisation of a Graded Word Reading Test*, p. 14.

the test as a whole gives an *average* pupil a mental age equal to his chronological age.

The lay public does not realise sufficiently the immense amount of work involved in calibrating a reliable mental scale of the Binet type. The latest revision of the Stanford-Binet scale required the full time of a team of research workers for nine years. The test-items were re-arranged and re-tested no fewer than six times in the course of the standardisation.

Standardising a Point-Scale.

The Binet scale was standardised directly in terms of units of mental age. In some cases however, e.g. in attainment tests in Arithmetic, English, Reading, Spelling, it is more convenient to assign marks (or points) for each correct answer. The total marks gained are arranged in age-groups e.g. 6-but-not-yet-7, etc. The average mark for each group is calculated. Then a graph is drawn on which the average mark scored by the 6-but-not-yet-7 group is made to correspond with an attainment age of VI years 6 months, and similarly for all the other age-groups covered by the test.* Then by means of the graph intermediate attainment ages can be read off corresponding to any given score.

Standardising a Test for a Restricted Age Range.

The use of tests of the Binet type revealed certain difficulties implicit in the definition of mental age, particularly at the upper limits of the scale and in the case of very bright pupils, for whom the test items are not sufficiently difficult. Thus, since the scale stops at a nominal average mental age of XVIII it is possible for very bright pupils of 14 years, bright pupils of 16 years and average pupils of 18 years, all to score the same mental age although they cannot be equal in mental capacity.

Again there is evidence which suggests that the *rate* of development of "intelligence," like that of height, slows down and finally ceases at some time after adolescence. Thus it is possible for an average adult of 18 years to score as high a mental age on the Binet scale as an average adult of 36 years. However, in terms of achievement-for-age we have no reason to suppose that the 18-year-old is twice as intelligent as the 36-year-old. They may quite easily

* See Schonell. *Diagnostic and Attainment Testing.* In actual practice the process is more complicated in detail. The scatter of the ages and marks at each age-grade must be checked by inspection of the standard deviations in order to estimate the possibility of any serious departures from a normal distribution which would suggest errors in selecting the population of pupils used for standardisation.

be equally intelligent. Thus, the Binet Scale ceases to give reliable indications at its upper age-limits.

In the second place, a test intended to cover a wide age-range can include only a relatively few test items at each age level. The latest revision of the Stanford-Binet scale includes 17 test items for year XI and 18 for year XII. Of these, 11 items must be passed successfully at each age level if the testee is to score the corresponding mental age. As against this, a Moray House group test of intelligence standardised for the chronological age-group consisting mainly of children between 10 years 6 months and 11 years 5 months includes 100 test items. Thus the restricted age-limit test is more sensitive than the Binet scale to very small differences in "intelligence" or attainment. The difference may be compared with that between one ruler graduated in half-inches and another in sixteenths.

The need for tests having greater sensitivity to small differences within a group of approximately the same chronological age has been emphasised by the demand for standardised tests for use in selecting pupils of age 11-plus for grammar schools. For this reason, and on account of the anomalies in the definition of mental age at the upper end of the Binet scale, psychologists have concentrated on standardising group tests for restricted age-limits. These tests require modifications in the methods of standardisation.

Omitting statistical technicalities which do not affect the main principles, the process goes as follows—

(i) more questions are assembled than will be needed for the final test. As many as four times the required number may be collected for an "intelligence" test.

(ii) the first draft is given to a trial sample of some 150 to 200 pupils adequately representative of the age-range in question.

(iii) the percentage of the group passing each item correctly, indicates as before, the relative difficulty of the item. This being known, all items answered correctly by less than 25% or more than 85% of the group are discarded.* It has been found by repeated experiment that better results are obtained by including more items of a moderate degree of difficulty at the expense of items which are very easy or very difficult.

(iv) the final draft is made up from the items now remaining in such a way that the *average* difficulty value of all the included items shall be 50%.

(v) as before, it is necessary that the test shall discriminate sufficiently between the different levels of ability and spread the candidates evenly along the scale. This is particularly important

* This is the effective answer to critics who object to the tests on the ground that no child at the age specified could possibly answer the questions. Actually, it has been found by trial that some pupils can, and do answer correctly.

if the test is to be used for purposes of selection for grammar schools. To measure power of discrimination, the test constructors proceed as follows.—The scripts are arranged in descending order of total marks. They are then divided into equal batches, six for example. For each test item, the number of pupils in each batch giving the correct answer is tabulated and set out as follows—

Batches arranged in descending order of scores	1 top	2	3	4	5	6 bottom
No. of pupils in each batch answering						
Test Item 1	20	18	10	8	3	1
Test Item 2	15	19	17	14	12	12
Test Item 3	—	—	—	—	—	—

In the case illustrated, almost as many pupils in the lowest batch have answered item 2 as in the top batch. This item, therefore, does not discriminate sufficiently between the bright and dull pupils.

By the use of a suitable formula, an index of discrimination for each item can be calculated. All items below a value found by experience to give the best results in practice are then discarded.

A sufficient number of the original test-items found on the first trial to be most satisfactory with respect to level of difficulty and power of discrimination are then given a second trial, this time on a much large sample—usually the whole 11-plus age-group in a county. Several thousand pupils may be used for this trial.

Great care must be taken to ensure that the final sample is *as nearly as possible representative, in due proportion, of all the various grades and conditions in the pupil population for which the test is intended.* From this representative sample, the final adjustments are made and the " mental scale " calculated. Suitable allowances are made for differences in chronological age within the limits prescribed by the test.

Finally, since the notions of mental age and intelligence quotient have now passed into general circulation, and are readily understood by teachers, the scores on these restricted-age-limit tests are transformed into " intelligence " quotients.*

Standardising the Instructions and Methods of Marking.

When the test items have been analysed and the scale calibrated the test constructors prescribe precise instructions for giving the test and marking the answers. This is to ensure that the tests will

* See later, p. 29.

be given and marked in the same way as that used in the process of standardisation itself. These instructions are issued in handbooks to accompany the tests. The appropriate handbook should always be studied carefully before a test is given since the accuracy of the results depends on the test being used in the same way *and in the same circumstances* as those in which it was originally standardised.

5

MENTAL AGES, ATTAINMENT AGES AND QUOTIENTS

Hitherto we have used the term "mental age" to describe the units in which the standardised scale is graduated. In practice, as was stated above, mental age is kept for scores on tests of "intelligence" or general educational capacity. These tests purport to measure innate mental capacity in contradiction to attainment tests which are measures of scholastic attainment in some particular subject. For the latter we use the term "attainment" age. The principles of standardisation are the same.

The mental or attainment age represents the level of mental development or scholastic achievement reached by a pupil of a given chronological age. From these two measures it is possible to compute a mental or attainment *quotient* (or ratio). This is done by dividing mental or attainment age by chronological age and multiplying by 100 (to remove fractions).
Thus—

$$\text{Intelligence Quotient} = \frac{\text{Mental Age}}{\text{Chronological Age}} \times 100$$

$$\text{Attainment Quotient} = \frac{\text{Attainment Age}}{\text{Chronological Age}} \times 100$$

The quotient is a measure of *rate* of development. If a 10-year-old pupil has a mental age of XII years his I.Q. is 120. That indicates that he has developed as far in 10 years as the normal or average pupil does in 12 years. A 10-year-old pupil with the mental age

of VII has an I.Q. of 70. He has developed in 10 years only as far as the average pupil of 7 years. Similarly for attainment quotients.

If the rate of mental development is constant it follows that the I.Q. or mental ratio might be used as an indication of probable future progress. Whether or not the rate of development and therefore the I.Q. is, in fact, constant is controversial. Most psychologists nowadays believe that it is approximately constant over a range of some two to three years, particularly during the primary school period i.e. 7 or 8 to 11 years. However predictions with respect to future mental development based on one test only must be accepted with caution as being no more than approximate. The mental age-scale, like all measuring instruments, is liable to errors of measurement. Also the course of mental development may be upset by radical changes of environment, emotional disturbances and other factors. For ordinary practical purposes we need to know whether a pupil's I.Q. falls below 90 or is within the ranges 90 to 110, 111 to 120, or above 120 since it is comparatively rare for a pupil's I.Q. to change from below 90 to above 110. If such a case is found it is most probable that some serious error was made in the original measurement.

Backwardness as Distinct from Educational Retardation.

As we noted above, a most important question nowadays is, to what extent is a pupil's attainment age commensurate with his mental age. In other words, is he working scholastically up to the level of his educational capacity. For example, a pupil of 11 may have a reading age of VIII years. However, if his mental age is only VIII years that pupil is working up to the limit of his capacity. He is backward *but not retarded*. On the other hand, if a pupil of 11 years has a mental age of XIII but, as sometimes happens, an arithmetic age of X years only, then that pupil is seriously retarded in Arithmetic but he is not backward in the sense of being dull. The distinction between retarded, and backward because dull, is most important in educational diagnosis and treatment and it can be made with confidence only on the evidence of well-standardised and reliable tests of "intelligence" and scholastic attainment. For diagnostic purposes the standardised tests are far superior to school marks and much more reliable than mere personal impressions.

6

WHAT DO STANDARDISED TESTS TEST?

VALIDITY AND RELIABILITY

Much of the suspicion with which standardised tests have been regarded by the general public has arisen from ignorance of the ways in which the tests have been constructed, and the purposes for which they are intended by their constructors.

From the previous section it should be clear that the tests indicate the standing of any given pupil relative to that of a hypothetical *average* pupil of his own chronological age. This follows from the way in which the tests are standardised, no matter whether they cover extended or restricted age-limits. Moreover, because the trial population in which the tests are standardised has been chosen deliberately and carefully to be representative of many classes, many schools, all grades of educable ability or scholastic attainment, each test is a sort of *common "yard-stick"* by which the work of different teachers and different schools can be fairly compared.

The disadvantage of the traditional non-standardised examinations was, precisely, that each teacher, or each inspector, set the questions according to his own personal subjective estimate of their difficulty for children of the type and age in question, and each one marked the answers according to his own methods and his own personal standards of value. Repeated surveys have shown beyond doubt that different examination papers intended for children of the same type and age have differed significantly in difficulty, and that different examiners' methods of marking and standards of mark value have varied enormously. Cases are reported in which an identical essay type of answer has been awarded all grades of merit from distinction to failure by different examiners all supposed to be competent for the purpose. The case of the non-standardised examination is very similar to that in which the yard measure was taken to be the length of any individual draper's arm. On the other hand the standardised test may be compared with the standard yard used by all drapers.

We have still to ask, however, to what extent a standardised test is *valid*. In other words, does it really test what it purports to test?

In the case of attainment tests this question is easily answered. A well constructed and standardised test of vocabulary, reading comprehension, English grammar or mechanical arithmetic is most likely to test, predominantly, attainment in vocabulary, reading comprehension, etc. A vocabulary test is not likely to measure

attainment in mechanical arithmetic or vice versa. It is comparatively easy to construct a valid test of attainment in some scholastic subject or other because we know without much doubt what it is we are attempting to measure.

On the other hand in the case of " intelligence " the problem is by no means simple. That is why, in this discussion, " intelligence " has been qualified by inverted commas.

The reason is, as in the case of many other terms in general use, e.g. " morality," " fair shares," " justice," that although " intelligence " has passed into common usage and although we imagine we know what it means, in actual fact nobody has succeeded yet in giving it an agreed general definition. This implies, of course, that the word is a " portmanteau " term which includes not one meaning but a complex set of different meanings not sorted out accurately in common usage.

It is in connection with the *validity* of intelligence tests that most of the controversies about standardised tests have arisen. What does an " intelligence " test actually test? We must find a reasonable answer to this puzzle. Otherwise we shall be in somewhat the same case as an individual who tries to measure pints of milk with a foot rule.

To establish the validity of an " intelligence " test we must first agree upon an *independent criterion* of intelligent behaviour and then compare the results of the tests with the criterion.

Consciously or otherwise, the independent criterion which has been used to check the validity of " intelligence " tests has been, in the last analysis, ability to succeed in school or college work as measured by the reports of competent observers familiar with the sample of pupils and students in question, together with final scholastic achievement.

For many years, follow-up surveys have been made in which a sufficiently large sample of pupils has been tested at a given age e.g. seven or eleven years. Their scholastic achievement and academic behaviour have been recorded over a period of years and the resulting order of merit compared with the original test results.

Because " intelligence " is so vague a term and because the really effective independent criterion of " intelligence " has been in fact subsequent progress in school or college, most of the commonly-called tests of general intelligence are really measures of educability (or general educational aptitude) and should be considered as such.

However, although the ultimate criterion for establishing the validity of " intelligence " tests has been educability, at the same time the analysis of test-items has clearly indicated which types of test-items are most closely correlated with, or saturated with " intelligence " as measured by the tests. Generally speaking, test-items which require the *application* of knowledge as distinct from

mere memorisation; reasoning; the appreciation of nice distinctions between the meanings of words; ingenuity, etc., are most highly saturated with whatever constitutes general "intelligence." It is significant that although those psychologists who have been most influential in devising tests of "intelligence" have differed widely in their theoretical definitions they have all used the same types of test-items for their practical tests.

It is important that this question of validity should be kept in mind. When some journalist interested mainly in causing a sensation picks out a particular test-item for ridicule he is apt to forget that *if the test in question has been reliably standardised*, all the test-items in it have been carefully analysed by experiment. It is not possible for anybody to judge by mere inspection, whether a given test-item is a valid test of educability. These matters can be settled only by an appeal to the results of practical experiment. In the historical development of tests it has been found that some test-items which were confidently expected to be highly saturated with "intelligence" had, in fact, little significant connection with it, while others which the test-constructors had at first regarded with suspicion were found to be highly indicative of educability.

These considerations need to be kept in mind when the lay public is invited to ridicule some test-item lifted from its context in some unspecified test.

Reliability.

A properly constructed test must not only be valid; it must also be *reliable*.

Reliability, in this context, means that if the same test is given to the same pupils on two or more occasions at intervals of time it will give closely similar results even if administered and marked by different people. Reliability is a measure of the trustworthiness of the test results.

Degree of reliability is indicated by a fraction. Complete reliability e.g. if a test gives *identical* results on two or more occasions —would be indicated by 1.0. A coefficient of reliability of 0.5 indicates that the test results are little, if any, better than guesswork. Test constructors publish reliability coefficients for standarised tests, and intending users should avoid any tests which have not a guaranteed high reliability coefficient. For routine school purposes reliability coefficients should be above 0.90 and tests for grammar school selection should have reliability coefficients not less than 0.95. To quote from a recent catalogue, four tests listed

therein, two of " intelligence " and two of attainment, have guaranteed reliability coefficients of 0.979; 0.952; 0.971; 0.976.*

7

CONCERNING THE CHOICE OF TESTS

We have now described how standardised tests are constructed; how the constructed tests are tested; and what, in fact, they measure.‡ It has seemed expedient to do this in some detail on account of the persistent and occasionally perverse criticisms and misunderstandings of these tests. Test-items are taken out of their context; no attention is paid to the way in which the items have been chosen, tested, and then combined within the test as a whole.

From these notes we can infer certain precautions which must be taken in choosing and using standardised tests. The following considerations must be kept in mind:

(a) only tests of a guaranteed high degree of reliability should be used;

(b) the tests must be given *strictly according to the instructions for administration and marking.* If a test is time-limited to a prescribed number of minutes then those limits should be accurately kept. If a stop-watch is prescribed then a stop-watch should be used.

(c) the tests should be used (i) only for the purpose for which they have been standardised (ii) only for the age-limits over which they have been standardised and (iii) only for pupils in *conditions similiar to those on whom the test was originally standardised.*

* Knowing the reliability coefficient of a test, statisticians can estimate the probable limits of accuracy of its results (provided, of course, that the test is given and marked strictly according to the prescribed instructions). For example, if a test has a reliability coefficient of 0.975 the chances are 68 in 100 that a quotient measured by it on a second occasion will not differ from the first by more than 2.37 points above or below; and 95 in 100 that the two results will not differ by more than 4.74 points above or below.

‡ The discussion has been restricted, for obvious reasons, to tests of educability and scholastic attainment. Numerous tests for specialised abilities are now available for use in vocational selection and guidance but the principles of construction and validation are the same as those described above.

No competent psychologist will claim that standardised tests are absolutely accurate measuring instruments. No competent psychologist will make a dogmatic prediction about *future* scholastic progress, on the basis of a single testing. Nevertheless, if used according to the principles stated above there is no doubt that the indications for educational guidance to be gained from a battery of well-chosen standardised tests are much more reliable than the results of an academic examination of the traditional type and much more useful for diagnostic purposes than impressions based on observations merely.

A lively controversy has arisen about the possible influence of practice in answering tests on the intelligence quotients of children, and about the effects of coaching.

There is no doubt that practice does raise test scores. It has been found however that increments of I.Q. due to practice in answering test-items diminish rapidly and that most pupils are "saturated" with practice after three or four trials. Some people seem to believe that this practice effect is a sufficient reason for abolishing the use of standardised tests. A more intelligent attitude would seem to be that all children should be tested at intervals during their primary school careers.

8

COMPARING RESULTS ON DIFFERENT TESTS. EFFECT ON QUOTIENTS OF METHODS OF STANDARDISING AND GRADUATING TESTS

As we have seen, the principle of a standardised scale of capacity or attainment is by no means new. Moreover, it is easy to understand and apply in practice, and the terms "mental age," "attainment age,' "intelligence quotient" have by now passed into general usage.

However, the use of standardised tests for school record purposes and for grammar school selection may lead to comparisons between I.Q's. or attainment quotients derived from different types of tests. It seems desirable, therefore, to point out at this stage that quotients are not absolute quantities. They are affected by the type of test used and by methods of standardisation.

This fact need not lead to any difficulty in practice provided that the purpose of the tests is clearly understood and that each test is used strictly for the purpose for which it was devised and standardised.

A Binet-type test is really a measure of intellectual or scholastic *development*. It indicates in units of years of mental or scholastic progress the level of development reached by the pupil tested when compared with the progress which a hypothetical average pupil in the same circumstances can be expected to make. The object of any extended age-range test is to reveal mental or scholastic age and by doing so to indicate how far in advance or in arrear is the pupil tested. Knowing this, the skilled teacher can estimate what exercises and methods are most likely to be educationally satisfactory for the pupil in question. For purposes of diagnosis, or of sorting out pupils within the same school for teaching purposes *it does not matter how many of the group may be approximately equal in development*. Information about the level of development, i.e. mental or scholastic age, is the most important consideration.

On the other hand, when all pupils between say 10½ years and 11½ years in an administrative county are to be tested in order to select a small proportion for admission to grammar schools, the authorities are not interested primarily in mental age and levels of development. They want to know which pupils in this 11-plus group are, intellectually and scholastically, the most able. In a group consisting of all the pupils in an administrative county, it is certain that a relatively large proportion will tend to cluster near the average mark of the group. Now, for the purposes of selection for grammar schools the administrators do not like the doubtful border-line cases where it is difficult to decide fairly, on objective evidence, between candidate A and candidate B. Therefore, in constructing and standardising a grammar-school selection test the psychologists aim deliberately at a form of test which will *separate out the candidates of equal chronological age as far as possible along the scale*. They are interested mainly in selection, not in educational guidance and remedial treatment.

Thus, mental or attainment age does not enter into the standardisation of a grammar-school selection test. Instead, each child in the group is assessed by measuring his standing in a representative group of pupils all exactly the same chronological age as himself. The degree of "intelligence" and attainment within the group is measured not in years of development but in units of the standard deviation of the distribution of the scores made by a representative group in the test. The same applies to all restricted-age-range tests.

This principle of measuring can be followed by reference to the data in Table I. The standard deviation of the distribution of marks of the 20 pupils exemplified there is 3.75; the average mark is 10. Consider pupil A. His mark is 17. His deviation above the average is 7 marks. In units of standard deviation this is $7 \div 3.75$, i.e. 1.87 units. Pupils with 14 marks will be just over 1 standard deviation unit above the group average. Pupils F and P are level with the group average. Their deviation from the average on this scale is 0.

Pupil C with 3 marks is 7 ÷ 3.75, i.e. 1.87 standard deviation units below the group average.

Thus it is possible to construct a scale graduated in units of standard deviation of the distribution of scores. In this way the relative standing in aptitude or attainment of individuals in a group of children all of the same chronological age can be measured and compared.

However, to the non-statistician, the notion of an intelligence quotient is more familiar than units of standard deviation. Therefore, for the greater convenience of education authorities the constructors of restricted-age-limit tests transform their standardised scores into "I.Q's". in such a way that the distribution of these I.Q's resembles that of the Binet-type tests. This can be done because, as we noted above, scores on "intelligence" and attainment tests made by large unbiassed samples of pupils tend to fall into a close approximation to normal distribution. Also, when the arithmetic mean and standard deviation of a normal distribution are known, the percentages of cases which may be expected to lie between the various intervals of the distribution can be calculated.

Thus, because the revised versions of the Binet scale give distributions of I.Q's with an average of 100 and standard deviation of approximately 16.5 the transformed I.Q's of the restricted-age-limit tests are roughly, but only *roughly* comparable with those of the latest revision of the Binet scale.* Inspection of the data in Table II will give some idea of the differences in the two I.Q. distributions.

This explanation of the effect on I.Q's and attainment quotients of different types of standardised tests and different methods of standardisation is necessary because it is often supposed, erroneously, that I.Q's are absolute quantities, and that the same child will score the same I.Q. no matter what type of test is used to measure it. Then people with this wrong idea may make illegitimate comparisons between I.Q's for the same pupil on two different types of tests and condemn the tests for not giving identical results. It would be just as absurd to cast doubts on the process of measuring height because the same man's height is 2 yards or 1.825 metres. The yards and metres represent scales graduated in different units.

For practical purposes the differences discussed above mean, not that standardised tests are unreliable but that in choosing tests sufficient care must be taken to select tests which have been standardised for the purpose for which they are required. Thus one would not use an extended-age-range test in order to select pupils of 11 years for grammar school entrance; or use a restricted-age-

* The average I.Q's in both scales will be 100. The effects of different units of measurement will increase as the I.Q's get farther away from the average, either way.

range test to establish the mental or attainment age, and, by implication, the degree of retardation of a pupil whose school work is unsatisfactory.

It follows from the above that additional precautions must be taken. They are (a) if it is necessary to *compare* the standing of several pupils for any purpose whatever then they must all have the same test; and (b) whenever a mental or attainment age, or quotient is recorded the name of the tests used to ascertain it should always be stated at the same time. Then there can be no more ambiguity about the statement than there is about the statement that X's height is 2 yards or 1.825 metres.

To make comparisons easier some test constructors nowadays are standardising all their tests on the same scale e.g. Moray House tests have a mean I.Q. of 100 and a standard deviation of 15.

9

PROBLEMS OF A BILINGUAL EDUCATIONAL POLICY FOR WALES

In 1953 in a report entitled "The Place of Welsh and English in the Schools of Wales" the Central Advisory Council for Education (Wales) recommended the adoption of a bilingual educational policy. By a bilingual policy the Council meant that the English-speaking population of Wales should acquire as satisfactory a control of the Welsh language as most Welsh-speaking children have of English.

This recommendation has been accepted by fourteen of the seventeen Local Education Authorities in Wales. The problem is, of course, how to implement it most efficiently in practice.

It would appear that standardised tests of attainment and educable capacity could be used with advantage in this connection.

1. For Survey Purposes.

If the policy of bilingual education is to be taken seriously, then periodic language surveys will be necessary to provide *reliable* estimates of the attainments of pupils of varying ages in Welsh and English languages. This is the only way of finding how much, if any, progress is being made.

Up to the time of writing official estimates of these attainments have been based on very dubious statistics. The Advisory Council's

own survey is a case in point. In stating the need for such a survey the Report says "The only attempts to make surveys in the past have been those of individual authorities in respect of their own schools . . . and taken as a whole they have suffered from the fact that there were no uniform criteria and no agreed standards or methods of interpreting the data not only as between one authority and another but even in the same authority at different periods." In other words, data from different areas and different observers are useless for comparative purposes unless estimated by means of a standard scale.

Unfortunately, the Council's own statistics suffer from precisely the same defect. All the teachers in Wales in 1950 were asked for the following information with respect to (i) pupils whose first language is Welsh; (ii) pupils whose first language is English: —

A. Number of children having no knowledge of the second language.

B. Number of children who can understand but not speak the second language.

C. Number of children who can understand simple lessons in the second language in such subjects as History, Geography or Nature Study and can conduct elementary conversations in the second language.

D. Number of children who can express themselves with fair fluency in the second language.

In a language survey conducted by the Welsh Joint Education Committee in 1961 the same criteria were used.

These criteria are, from a statistical point of view, shockingly vague. For example, what is meant, exactly, by no knowledge? What is meant by understanding but not speaking the second language? How much understanding is necessary to entitle a particular pupil to inclusion? What is meant by "understanding simple lessons"? How simple must a lesson be for this purpose? At what age is understanding 'simple' in this connection; for example, is 'simple' at age seven equivalent to 'simple' at age eleven? What is 'elementary' conversation at any given age level? What is a 'fair fluency' and is any fair fluency in Welsh equivalent to a fair fluency in English?

Vague standards such as these will be decided by the subjective estimates of men and women teachers of different ages in different grades and different types of schools; teachers moreover with varying language backgrounds and levels of competence in Welsh and English from high efficiency to little or none at all. Again, attitudes toward Welsh or English will constitute distorting factors in judgment. Those anxious to make a good show for one or other language will include as many as can be crowded into categories C. and D. Those who are openly or secretly opposed to the bilingual policy will be more likely to believe that the policy is undesirable

or impossibly 'idealistic' and include as many as possible in categories A and B. Moreover, there is evidence for supposing that people show a tendency to be either predominantly 'includers' or 'excluders' when they are required to classify and this constitutes a temperamental difference of which the classifiers are quite unaware.

It will be suggested, of course, that by and large these errors of judgment will cancel each other out. In the case of purely *chance* errors that may be correct. However, errors of judgment arising from the sources we have indicated above are not chance errors. They are systematic errors and they will be cancelled out only if as many people with positive sets of attitudes are equal in numbers, and prestige, and intensity of attitude to those with negative sets. Concerning this important qualification there is no evidence whatever. In addition, cancelling out differences in the process of averaging may hide differences which are vital to an adequate understanding of the situation in question. We still await reliable data with respect to the distribution of English and Welsh speech both geographically, and in terms of grades of educable capacity, attainment and linguistic background. Such data can be provided by standardised attainment tests in vocabulary, oral reading and comprehension.

2. For Estimating Degrees of Educable Capacity.

To implement the bilingual policy the schools must teach both languages. This raises the question of educable capacity in relation to learning two languages simultaneously and teaching them. The Report states the problem thus: "Having due regard to the varied abilities and aptitudes of the pupils and to the varied linguistic patterns in which they live, the children of the whole of Wales and Monmouthshire should be taught Welsh and English *according to their ability to profit by such instruction.*[1] This recommendation implies that consideration should be given to the desirability of teaching only their mother-tongue to children with relevant physical or mental disabilities and to children whose lack of ability together with a poor supporting linguistic background makes the learning of a second language whether English or Welsh acutely burdensome."

This puts some obvious difficulties of a bilingual policy quite bluntly. One may ask then, how the relevant disabilities can be diagnosed, and measured and by whom? At what level, precisely, of mental capacity and linguistic ability shall be the line be drawn, and who will decide which pupils shall not be taught a second language?

The Council suggests that to impose a uniform aim on all pupils must be undesirable for at least two reasons: —

[1] Italics mine.

(a) the standards would be too low for the abler pupils; this would lead to boredom and frustration.
(b) the standards would be too high for the weaker pupils resulting in failure and loss of confidence with the consequent aversion from further learning. The main consideration must be the relation between a pupil's aptitude, ability and linguistic background, and the level of attainment to be expected. As for who shall decide these matters, the answer proposed is, the teachers. As usual, they are left holding the baby.

How are the teachers to decide? To implement the Advisory Council's suggestions, they must discover in the case of each pupil:

(i) General educable capacity;
(ii) linguistic capacity (not by any means identical with (i);
(iii) present level of attainment in English and Welsh;
(iv) the linguistic background in the neighbourhood;
(v) what levels of attainment and rate of learning are appropriate having due regard to items (i) and (iv) above.

Again, for greater efficiency it is desirable to decide (1) which, if any, of several alternative methods of teaching the first and second language is most effective for pupils of a given level of linguistic and general capacity and attainment; (2) which is the best way of organising schools and classes in cases where a selective examination is used for grammar-school entrance; (3) what level of attainment is to be expected or demanded from pupils with different levels of capacity and ability and different linguistic backgrounds.

It would appear obvious, therefore, that if the bilingual policy is to be administratively and educationally efficient, tests of educable capacity and linguistic attainment are needed which have been correctly standardised on sufficiently large and representative populations of children in Wales *according to their linguistic background* (unless, of course, everybody prefers to muddle along as usual on platitudes and sentiment).

This may appear to be a formidable task. However, for purposes of guidance to administrators and teachers it would be sufficient to conduct periodical tests in sample schools in representative areas varying from predominantly Welsh to predominantly English, both urban and rural.[1]

This forces upon our attention certain special problems involved in constructing and using standardised tests in a mixed language area such as Wales is at present.

[1] See, for example, the survey undertaken at the request of the Welsh Joint Education Committee by W. R. Jones, J. R. Morrison, J. Rogers, H. Saer on "The Educational Attainment of Bilingual Children in Relation to their Intelligence and Linguistic Background." University of Wales Press 1957. Relevant to our discussion is the authors' warning that their findings must be viewed within the limits inevitably imposed on their work *by the lack of suitable and satisfactory test material in Welsh.*

10

CONSTRUCTING AND USING STANDARDISED TESTS IN A MIXED LANGUAGE AREA

In Wales the linguistic background is mixed in various degrees from almost or quite monoglot English to almost monoglot Welsh. This raises problems in the production and use of standardised tests of which educators and investigators have not always been explicitly aware.

A glance through the earlier published accounts of investigations into bilingualism will show that the experiments were made on *two supposedly homogeneous linguistic groups*—one monoglot English, the other 'bilingual.' But no attempts were made to discover *to what extent* the supposedly bilingual group was, in fact, bilingual. Generalisations about the effects of bilingualism or about methods of dealing with supposedly bilingual groups based upon these experiments could be quite misleading. Moreover until the researches of W. R. Jones on the possible influences of socio-economic status it was not realised that such status was correlated positively with linguistic background and educational attainment.[1] In any discussions about bilingualism in relation to educable capacity and scholastic attainment, both linguistic background and socio-economic status of the pupils need to be estimated and allowed for.

In constructing and using standardised tests in Wales two fundamental principles must always be kept clearly in mind. They are: —

(a) Standardised tests can only be used, reliably, on populations *equivalent to that on which the tests were originally standardised.*

(b) the population used for purposes of standardisation should vary significantly *only with respect to the variable which that test is supposed to measure*, e.g. educable capacity, or linguistic aptitude or linguistic attainment.

With regard to (a) from the account of the standardising process given above it is obvious that a test standardised on an English population may be not at all suitable for use in Wales, not even for testing monoglot English-speaking pupils. Its suitability and

[1] W. R. Jones. "Bilingualism and Intelligence." University of Wales Press, 1959. The author states
"It appears, for example, that occupations in the professional, salaried and non-manual categories predominate in the English and mixed-English groups whereas, by contrast, the observed frequencies of such occupations fall distinctly below the expected frequencies in the case of the Welsh group." p. 42.

validity cannot be taken for granted until the case has been proved by a re-standardisation which will check the achievement-for-age norms of the population in question.

Mere translations of standardised English tests into Welsh will not give reliable results until an adequate item-analysis has been made on a representative Welsh group. There is no guarantee whatever that the difficulty values and discrimination values of the Welsh translations will be equivalent to those of the original English versions. Neither difficulty nor discrimination values can be estimated by inspection. They can be found only only by actual trial on a representative population. Moreover, tests standardised on predominantly urban populations particularly in Wales will not give reliable norms if used subsequently on rural populations. The background conditions are not the same in the two cases.

With regard to principle (b) above, the reason is not, at first sight, so obvious. Suppose we wish to standardise a test in problem arithmetic. The pupils selected for the purpose must be capable of *reading the words* in which the problems are stated sufficiently well to understand clearly what the problems are about, and sufficiently rapidly to be able to complete a time-limited test within the limits allowed.[1] If they cannot do so then even if they are familiar with the appropriate mechanical calculations they cannot solve the problems because they will not understand clearly what the problems are. Unless there is in the tested group a minimum ability to read which is approximately the same for all members of the group the test will measure an indeterminate mixture of attainments in both reading comprehension and arithmetic. It will measure neither accurately.

Similarly in the construction and use of tests of "intelligence."

It has been demonstrated again and again that success on a standardised verbal "intelligence" test depends on linguistic ability, on social status and home background. This is not difficult to understand. Whatever "intelligence" may be, it can only reveal itself through the medium of acquired experience. If a child is deprived of the common experience of its native culture then whatever innate powers may be involved in intelligent behaviour will find only inadequate means of expression. The items in the original Binet tests of general intelligence were chosen and worded on the supposition that they could be answered by means of activities and experiences which a child in that culture background might have been expected to "pick up" in the ordinary course of living at home, or in school, or at play. In other words, it was taken for granted that in every respect except the aptitudes which are involved

[1] If, indeed, there is to be a time limit. It has been shown that rural children are seriously handicapped in time-limited tests. See Morton and Butcher, Brit. Jnl. Ed. Psych. XXXIII, 1 Feb, 1963, p. 22.

in intelligent behaviour all the children to be tested started equal. This supposition must not be taken for granted.

Before any standardised test can be used reliably for purposes of comparison there must be some guarantee that all the individuals in the population tested have had *equal opportunities* both through in-school teaching and out-of-school experiehce to develop the aptitudes or acquire the attainments which it is the purpose of the test to detect and measure.

Thus, in testing bilingual populations in a mixed language area some means of estimating degrees of bilingual background and socio-economic status must be available both during the processes of standardisation and subsequently in testing bilingual populations. It has been shown, for example in the construction of a "Welsh Linguistic Background Scale"[1] that there was in a mixed language population used for the investigation a coefficient of correlation between Welsh Linguistic Background scores and scores on a Welsh Vocabulary Test of 0.85 which indicates a close relation between the two; the richer the Welsh background of the pupil the higher the score on the Vocabulary test.

These results have an obvious bearing on the problem of constructing standardised tests in Welsh, in a mixed language area.

In the first place, the population is not homogeneous with respect to linguistic background and attainment in Welsh. Not all the pupils have equal opportunities for acquiring either Welsh or English speech. Differences in linguistic background " saturation " must be estimated and allowed for in determining the norms of attainment which may reasonably be expected of any given " bilingual " pupil.

Put in another form the question is—What is, in fact, a representative sample of Welsh pupils with respect to attainment in Welsh or English? Suppose, for example, we wish to construct and standardise a test of educable capacity (" intelligence ") verbal or non-verbal, or of attainment in some subject involving language. What sort of sample should be used for estimating the attainment-levels which may be expected of a hypothetically average child at a given chronological age? If a sample is selected containing all grades of language mixture from mainly Welsh to mainly English, then those norms must not be used, strictly speaking, to test the attainments of other samples unless the *proportions* of language mixture are comparable. The same principle applies to socio-economic status. Moreover, the norms so obtained would be useless for practical guidance. As an indication of the levels of attainment-for-age and rates of progress which may reasonably be expected from pupils learning English and Welsh concurrently, norms based on the *average* of all degrees of language mixture will be misleading since they will effectively mask the varying influence of different degrees

[1] M. E. Gwenda Rees. "A Welsh Linguistic Background Scale." Aberystwyth Collegiate Faculty of Education. Pamphlet No. 2. 1954.

of Welsh and English background saturation and of socio-economic status on attainment in either language. To set the same targets of attainment-for-age in Welsh for pupils with all degrees of Welsh linguistic background will be absurd, as the Advisory Council said in its report. Norms for the standardised tests must be computed for various ranges of linguistic background scores. Further, when tests standardised in this way are to be used to measure educable capacity or scholastic attainment the linguistic background scores of the pupils to be tested must first be ascertained.

These problems are not insoluble. What it amounts to is that a good deal of preliminary research needs to be done by competent research workers to provide the basic data required for further investigations. A start has already been made. The following tests in Welsh are already available: —

REES, M. E. GWENDA. *A Welsh Linguistic Background Scale.* Collegiate Faculty of Education, Aberystwyth. Pamphlet No. 2. 1954.

A Linguistic Background Scale
prepared by members of the Collegiate Faculty of Education, University College of North Wales, Bangor. See T. R. MILES "Bilingualism in Caernarvonshire" published by the University of Wales Press.

A Linguistic Background Questionnaire (1960)
prepared by W. R. JONES and J. R. MORRISON for the National Foundation for Educational Research. Form N.S. 74 B.

Profion Dealltwriaeth Di-iaith.
A Welsh adaptation by W. R. JONES of Jenkins' Scale of Non-Verbal Mental Ability (with a Manual of Instruction in Welsh), Age-range 10-12, published by the National Foundation for Educational Research for the Collegiate Faculty of Education, University College of North Wales, Bangor.

Cotswold Mental Ability Test (Verbal) for the age-range 10-12 years. Adapted into Welsh by W. R. JONES from the original test by J. W. Jenkins, published by R. Gibson & Sons, Queen St., Glasgow, C.1.

BRACE, J. L. *A Welsh Word-Recognition Test.* Collegiate Faculty of Education, Aberystwyth. Pamphlet No. 5.

EMMETT, W. G. *Dee-side Non-Verbal Reasoning Test No.* 1 for age-range 10-12 years.
This is a non-verbal test in which the instructions are given in both English and Welsh prepared for the Collegiate Faculty of Education, Aberystwyth, in co-operation with the Carmarthenshire L.E.A. Published by Harrap & Co. Ltd., High Holborn, London, W.C.1. (A closed test for use by Local Education Authorities only).

Prawf Cymraeg (CI). An attainment test in Welsh Language Usage intended for pupils aged 10 years 5 months to 11 years 6 months having Welsh as their first language. (Available to Local Education Authorities only). Published by the Collegiate Faculty of Education, University College of Wales, Aberystwyth.

BIBLIOGRAPHY

1 General Introductions to the History, Theory and Practice of Mental Testing.

Knight, R.	*Intelligence and Intelligence Testing.* Short, interesting, easily followed.	Methuen. 1950
Ballard, P. B.	*Mental Tests.* A clear introduction to the subject with notes on elementary statistics.	University of London Press 1949.
Ballard, P. B.	*The New Examiner.* A critique of the easy-type examination.	University of London Press. 1949.
Freeman F. N.	*Mental Tests. Their History, Principles and Application.* An account of the historical development of mental testing. Discussions about the theory and techniques of standardisation. Notes on the use of tests and the interpretation of results.	Harrap. 1939.
Vernon, P. E.	*Intelligence and Attainment Tests.* Discusses the underlying principles of test construction and gives a clear critical account of the interpretation of test results. Does not need any special knowledge of statistics. A useful introduction.	University of London Press. 1960.
Watts, A. F.	*Can we Measure Ability?* A short account of testing written with special reference to some of the questions most commonly asked by parents and teachers.	University of London Press. 1953.

2. Practical Applications of Testing in Schools.

Glassey, W., and Weeks, E. J.	*The Educational Development of Children.* A teacher's guide to the keeping of school records. Contains chapters on "intelligence" tests, tests of attainment, tests for special abilities and interests, elementary statistics. Includes examples of record cards and notes on details of available standardised tests. A useful, practical introduction.	University of London Press. 1950.
Greenhough, A. and Crofts, F. A.	*Theory and Practice in the New Secondary Schools.* A review of problems in secondary education subsequent to the 1944 Education Act. Includes a chapter on simple statistics and testing.	University of London Press. 1949.

3. The Binet Scale.

Burt, C. L.	*Mental and Scholastic Tests.* A comprehensive account of the standardisation of an English version of the Binet Scale. Deals with the theory and methods of standardisation. Contains also a memorandum on tests of educational attainment and their uses.	Staples Press. 1949.
Terman, L. M., and Merrill, M. A.	*Measuring Intelligence.* A guide to the administration of the new revised Stanford-Binet test of intelligence. Contains the complete revised Stanford-Binet Scale with detailed instructions for administration and marking.	Harrap. 1949.

4. Details of Tests of Various Types and Their Practical Applications.

Vernon, P. E.	*The Standardisation of a Graded Word Reading Test.* Describes the standardisation of an attainment test in reading, and directions for applying it.	University of London Press.

VERNON. P. E.	*The Measurement of Abilities.* A critical account of the theory and practice of testing. Includes details of tests for various purposes; hints to testers; notes on new-type examinations; notes on statistical methods. Bibliography.	University of London Press. 2nd Edition. 1956.
HUNT, E. P. A., and SMITH, P.	*A Guide to Intelligence and Other Psychological Testing.* A short introduction to the subject with particular reference to the use of standardised tests in vocational guidance and selection.	Evans Bros. 1951.
BALLARD, P. B.	*Group Tests of Intelligence.* Contains examples of group tests with hints on their application. Notes on the nature of intelligence and on elementary statistics.	University of London Press. 1953.
CATTELL, R. B.	*A Guide to Mental Testing.* A comprehensive reference book. Notes on the nature and use of a large number of standardised tests for various purposes. Hints on selection of tests, interpretation of results, case studies, statistical formulæ.	University of London Press. 1953.
BURT. C. L.	*A Handbook of Tests.* A selection of Burt's standardised tests collected in one book for easy reference.	Staples Press.
SCHONELL, F. J., and SCHONELL, F. E.	*Diagnostic and Attainment Testing.* Contains all the Schonell tests of attainment and diagnostic tests in Reading, Spelling, Arithmetic, English Language in one volume for easy reference, together with notes on standardisation and the use and interpretation of the tests.	Oliver and Boyd. 1950.

5. BACKWARDNESS, EDUCATIONAL RETARDATION AND REMEDIAL METHODS.

BURT, C. L.	*The Backward Child.* Contains chapters on testing and classification.	University of London Press.
SCHONELL, F. J.	*Backwardness in the Basic Subjects.* A standard book on diagnosis and remedial treatment.	Oliver and Boyd.

SCHONELL, F. J.	*Diagnosis of Individual Difficulties in Arithmetic.*	Oliver and Boyd.
SCHONELL, F. J.	*Psychology and Teaching of Reading.*	Oliver and Boyd.
SCHONELL, F. J.	*Essentials in Teaching and Testing Spelling.*	MacMillan.

6. TRANSFORMING SCHOOL MARKS AND ESTIMATES TO A STANDARD SCALE.

PAMPHLET	*The Standardisation of School Marks.* A simple account of some short methods of transforming school marks to a standard scale.	University of Nottingham Institute of Education
MCINTOSH, D. M. WALKER, D. A., and MACKAY, D.	*The Scaling of Teachers' Marks and Estimates.* A more advanced treatment of standard scales and methods of transformation. Requires some knowledge of statistics.	Oliver and Boyd. 1949.

7. STATISTICAL METHODS.

ELDERTON, W. P. and ELDERTON, E. M.	*A Primer of Statistics.* A very simple and interesting introduction to the fundamental concepts underlying statistical methods. An excellent first book.	A. and C. Black.
MORONEY, M. J.	*Facts and Figures.* A layman's introduction to statistics.	Pelican Series.
LEVY, H., and PREIDEL, E. E.	*Elementary Statistics.* A good introduction for readers with some knowledge of mathematics.	Nelson. 1945.
GARRETT, H. E.	*Statistics in Psychology and Education.* A comprehensive exposition of statistical methods with special reference to educational problems. Requires only a working knowledge of arithmetic.	Longmans. Green & Co.

CHAMBERS, E. G.	*Statistical Calculations for Beginners.* A very useful book of reference. It explains as simply as possible how to perform the calculations involved in the commoner statistical methods. The calculations described involve a knowledge of arithmetic only. Worked examples and exercises included.	Cambridge University Press. 1948.
DAWSON. S.	*An Introduction to the Computation of Statistics.* Another good introduction to methods of statistical calculation. This one needs a rather more advanced mathematical background than does Chambers. Includes worked examples and exercises.	University of London Press. 1933.